הכשרות AI
NLP
תרגיל בית 2

אייל שוורצמן
ת.ז. 033575077

המשימה

בתרגיל זה התבקשנו לאמן מודל שמסווג ביקורות על סרטים

מאגר הביקורות מכיל כ 50,000 ביקורות והוא מתפצל בצורה מאוזנת בין ביקורות שליליות לחיוביות.

בחירת הקידוד

לטובת האימון, נבחנו 2 מודלים:
- מודל Twitter, קרוב יותר לדומיין של הסקירות (טקסט חופשי של גולשים באינטרנט)
- מודל Common Crawl 840B, מספר מילים גדול הרבה יותר

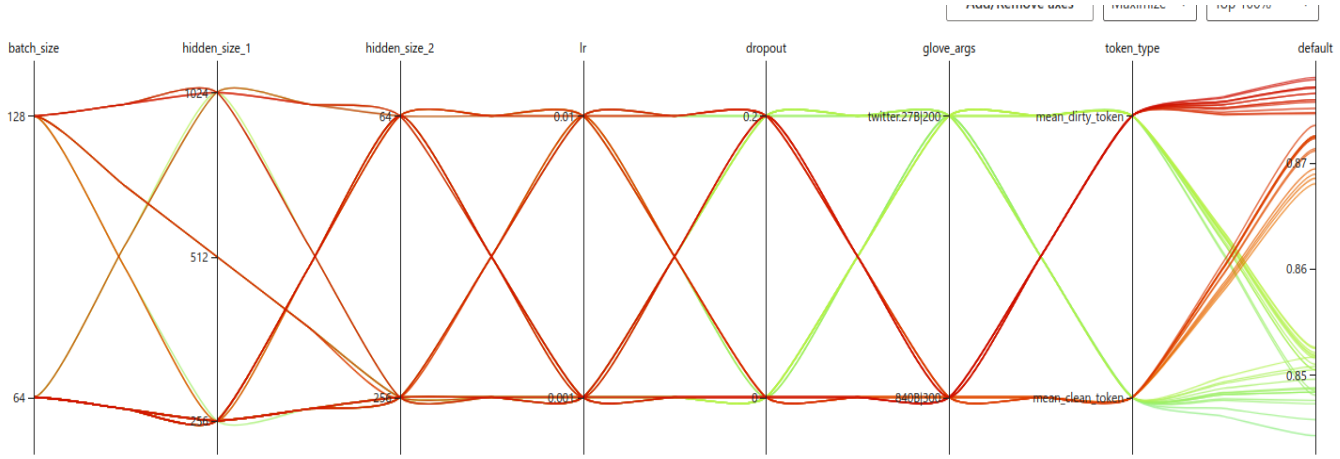
ישנה שונות גדולה בין מספר המילים בכל ביקורת, העדפתי לחשב את הממוצע, בדומה לתרגיל כיתה
השוויתי גם (בהמשך) בין סיווג על מיצוע של טקסט גולמי (mean_clean_token) למיצוע של טקסט נקי (mean_dirty_token)

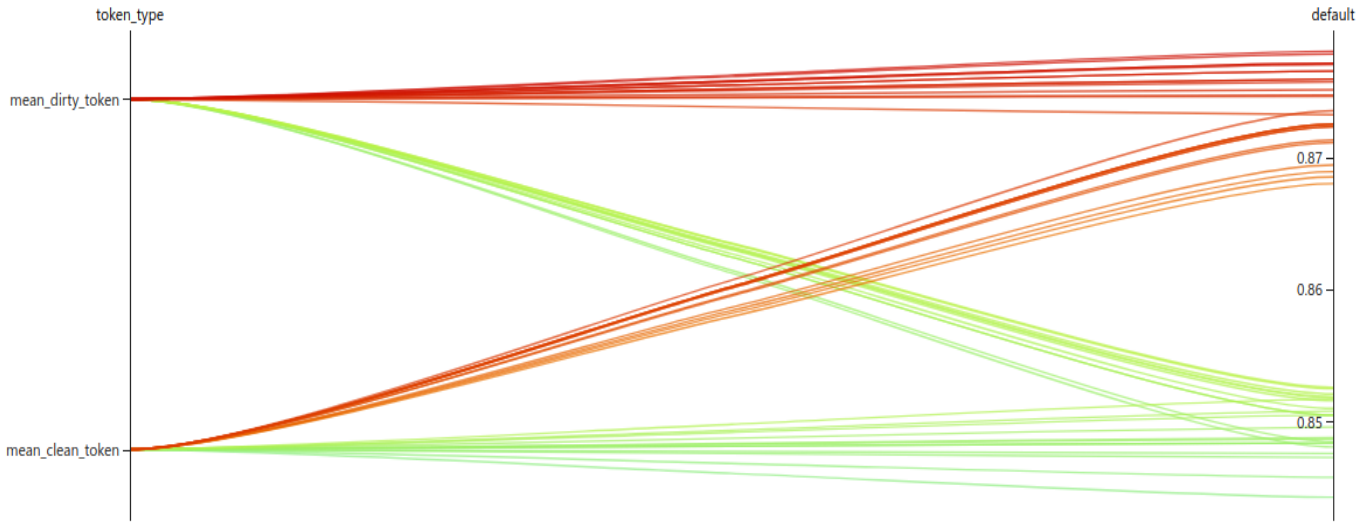
בניית רשת נוירונים

את רשת הנוירונים בניתי בהתאם למה שהודגם בתרגיל כיתה, רשת אשר מכילה שרשור של 2 שכבות FC אשר בין כל שכבה לשכבה אקטיבציה לא לינארית של RELU.

בחירת היפר פרמטרים

שימוש ב microsoft NNI על מנת לבחור את ההיפר פרמטרים, כאשר אופטימזציית המטרה היא accuracy:





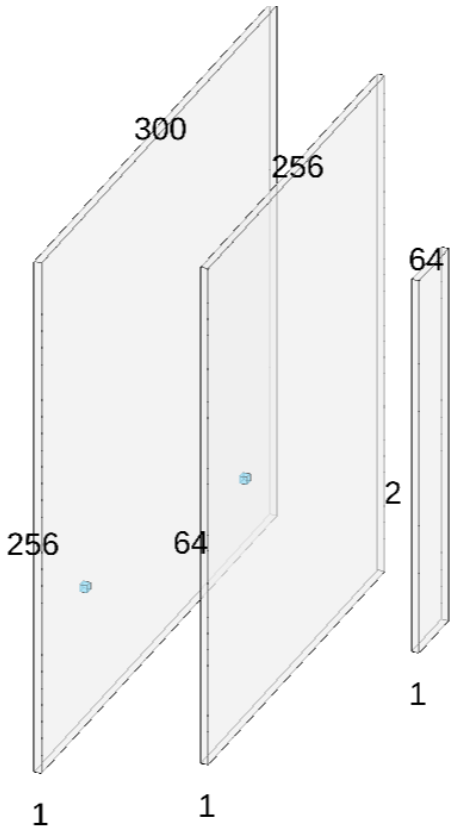
batch_size (space)	hidden_size_1 (space)	hidden_size_2 (space)	lr (space)	glove_args (space)	token_type (space)	Default metric
64	256	64	0.001	840B|300	mean_dirty_token	0.878 (FINAL)
128	1024	256	0.001	840B|300	mean_dirty_token	0.8778 (FINAL)
64	256	256	0.001	840B|300	mean_dirty_token	0.8771 (FINAL)
128	512	256	0.001	840B|300	mean_dirty_token	0.8771 (FINAL)
64	256	64	0.01	840B|300	mean_dirty_token	0.877 (FINAL)
64	256	256	0.01	840B|300	mean_dirty_token	0.8765 (FINAL)
128	1024	64	0.001	840B|300	mean_dirty_token	0.8765 (FINAL)

כפי שניתן לראות ניקוי הטקסט לא תרם לעלייה בדיוק, שאר הפרמטרים:

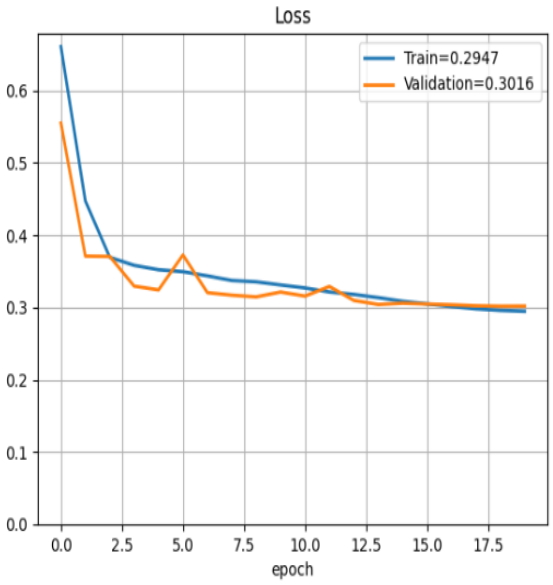
Batch Size:	64
Hidden Size 1:	256
Hidden Size 2:	64
LR:	0.001
Glove:	840B | 200

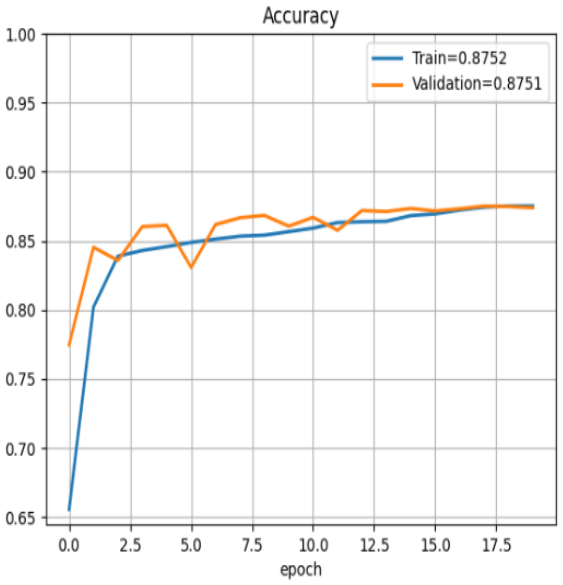
מבנה הרשת

אתר



תוצאות


Loss
Train=0.2947
Validation=0.3016
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
epoch


Accuracy
Train=0.8752
Validation=0.8751
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
epoch